Discourse Annotation of Non-native Spontaneous Spoken Responses Using the Rhetorical Structure Theory Framework

Xinhao Wang[1], James V. Bruno[2], Hillary R. Molloy[1], Keelan Evanini[2], Klaus Zechner[2]
Educational Testing Service
[1]90 New Montgomery St #1500, San Francisco, CA 94105, USA
[2]660 Rosedale Road, Princeton, NJ 08541, USA
xwang002, jbruno, hmolloy, kevanini, kzechner@ets.org

Abstract

The availability of the Rhetorical Structure Theory (RST) Discourse Treebank has spurred substantial research into discourse analysis of written texts; however, limited research has been conducted to date on RST annotation and parsing of spoken language, in particular, non-native spontaneous speech. Considering that the measurement of discourse coherence is typically a key metric in human scoring rubrics for assessments of spoken language, we initiated a research effort to obtain RST annotations of a large number of non-native spoken responses from a standardized assessment of academic English proficiency. The resulting inter-annotator κ agreements on the three different levels of Span, Nuclearity, and Relation are 0.848, 0.766, and 0.653, respectively. Furthermore, a set of features was explored to evaluate the discourse structure of non-native spontaneous speech based on these annotations; the highest performing feature showed a correlation of 0.612 with scores of discourse coherence provided by expert human raters.

1 Introduction

The spread of English as the global language of education and commerce is continuing, and there is a strong interest in developing assessment systems that can automatically score spontaneous speech from non-native speakers with the goals of reducing the burden on human raters, improving reliability, and generating feedback that can be used by language learners. Discourse coherence, which refers to the conceptual relations between different units within a response, is an important aspect of communicative competence, as is reflected in human scoring rubrics for assessments of non-native English (ETS, 2012). However, discourse-level features have rarely been investigated in the context of automated speech scoring. This study aims to construct a discourse-level annotation of non-native spontaneous speech in the framework of Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), which can then be used in automated discourse analysis and coherence measurement for non-native spoken responses, thereby improving the validity of the automated scoring systems.

RST is a descriptive framework that has been widely used in the analysis of discourse organization of written texts (Taboada and Mann, 2006b), and has been applied to various natural language processing tasks, including language generation, text summarization, and machine translation (Taboada and Mann, 2006a). In particular, the availability of RST annotations on a selection of 385 Wall Street Journal articles from the Penn Treebank[1] (Carlson et al., 2001) has facilitated RST-based discourse analysis of written texts, since it provides a standard benchmark for comparing the performance of different techniques for document-level discourse parsing (Joty et al., 2013; Feng and Hirst, 2014).

Another important application of RST closely related to our research is the automated evaluation of discourse in student essays. For example, one study used features for each sentence in an essay to reflect the status of its parent node as well as its rhetorical relation based on automatically parsed RST trees with the goal of providing feedback to students about the discourse structure in the essay (Burstein et al., 2003). Another study compared

[1] https://catalog.ldc.upenn.edu/LDC2002T07

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 263–268
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-2041

features derived from deep hierarchical discourse relations based on RST parsing and features derived from shallow discourse relations based on Penn Discourse Treebank (PDTB) (Prasad et al., 2008) parsing in the task of essay scoring and demonstrated the effectiveness of deep discourse structure in better differentiation of text coherence (Feng et al., 2014).

Related work has also been conducted to annotate discourse relations in spoken language, which is produced and processed differently from written texts (Rehbein et al., 2016), and often lacks explicit discourse connectives that are more frequent in written language. Instead of the rooted-tree structure that is employed in RST, the annotation scheme with shallow discourse structure and relations from the PDTB (Prasad et al., 2008) has been generally used for spoken language (Demirahin and Zeyrek, 2014; Stoyanchev and Bangalore, 2015). For example, Tonelli et al. adapted the PDTB annotation scheme to annotate discourse relations in spontaneous conversations in Italian (Tonelli et al., 2010) and Rehbein et al. compared two frameworks, PDTB and CCR (Cognitive approach to Coherence Relations) (Sanders et al., 1992), for the annotation of discourse relations in spoken language (Rehbein et al., 2016).

In contrast to these previous studies, this study focuses on monologic spoken responses produced by non-native speakers within the context of a language proficiency assessment. A discourse annotation scheme based on the RST framework was selected due to the fact that it can effectively demonstrate the deep hierarchical discourse structure across an entire response, rather than focusing on the local coherence of adjacent units.

2 Data

This study obtained manual RST annotations on a corpus of 600 spoken responses drawn from a large-scale, high-stakes standardized assessment of English for non-native speakers, the TOEFL® Internet-based Test (TOEFL® iBT), which assesses English communication skills for academic purposes (ETS, 2012). The speaking section of the TOEFL iBT assessment contains six tasks, each of which requires the test taker to provide an unscripted spoken response 45 or 60 seconds in duration. The corpus used in this study includes 100 responses from each of six different test questions that comprise two different speaking tasks: 1) Independent questions: providing an opinion based on personal experience (N = 200 responses) and 2) Integrated questions: summarizing or discussing material provided in a reading and/or listening passage (N = 400 responses). The spoken responses were all manually transcribed using standard punctuation and capitalization. The average number of words per response is 104.4 (st. dev. = 34.4) and the average number of sentences is 5.5 (st. dev. = 2.1).

The spoken responses were all provided with holistic English proficiency scores on a scale of 1 to 4 by expert human raters in the context of operational, high-stakes scoring for the spoken language assessment. The scoring rubrics address the following three main aspects of speaking proficiency: delivery (pronunciation, fluency, prosody), language use (grammar and lexical choice), and topic development (content and coherence). In order to ensure a sufficient quantity of responses from each proficiency level, 25 responses were selected randomly from each of the 4 score points for each of the 6 test questions.

The current study builds on a previous study that investigated approaches for modeling discourse coherence in non-native spontaneous speech (but which did not consider the hierarchical rhetorical structure of speech) (Wang et al., 2013). In that study, each spoken response in the same corpus that was used for the current study was provided with global discourse coherence scores. Two expert annotators (not drawn from the pool of expert human raters who provided the holistic scores) provided each response with a score on a scale of 1 to 3 based on the orthographic transcriptions of the spoken response. The three score points were defined as follows: 3 = highly coherent (contains no instances of confusing arguments or examples), 2 = somewhat coherent (contains some awkward points in which the speaker's line of argument is unclear), 1 = barely coherent (the entire response was confusing and hard to follow). In addition, the annotators were specifically required to ignore disfluencies and grammatical errors as much as possible. The inter-annotator agreement for these coherence scores was κ = 0.68. These discourse coherence scores are reused in the current study (along with the holistic profiency scores presented above) to evaluate the performance of features measuring discourse coherence based on the RST annotations.

3 Annotation

3.1 Guidelines

We used a modified version of the tagging reference manual for the RST Discourse Treebank (Carlson and Marcu, 2001) for this study. According to these guidelines, annotators segment a transcribed spoken response into Elementary Discourse Unit (EDU) spans of text (corresponding to clauses or clause-like units), and indicate rhetorical relations between non-overlapping spans which typically consist of a nucleus (the most essential information in the rhetorical relation) and a satellite (supporting or background information). In contrast to well-formed written text, non-native spontaneous speech frequently contains ungrammatical sentences, disfluencies, fillers, hesitations, false starts, and unfinished utterances. In some cases, these spoken responses do not constitute coherent, well-formed discourse. On the other hand, spoken responses are relatively shorter and comprise simpler discourse structures with fewer relations, which simplifies the RST annotation task in comparison to written text. In order to account for these differences, we created an addendum to the RST Discourse Treebank manual introducing the following additional relations: Disfluency relations (in which the disfluent span is the satellite and the corresponding fluent span is the nucleus), Awkward relations (corresponding to portions of the response where the speaker's discourse structure is infelicitous; awkward relations are based on pre-existing relations, such as *awkward-Reason*, if the intended relation is clear but is expressed incoherently or are labeled as *awkward-Other* if there is no clear relation between the awkward EDU and the surrounding discourse), Unfinished Utterance relations (representing EDUs at the end of a response that are incomplete because the test taker ran out of time in which the incomplete span is the satellite and the root node of the discourse tree is the nucleus), and Discourse Particle relations (such as *you know* and *right*, which are satellites of adjacent spans).

The discourse annotation tool used in the RST Discourse Treebank[2] was also adopted for this study. Using this tool, annotators incrementally build hierarchical discourse trees in which the leaves are the EDUs and the internal nodes correspond to contiguous spans of text. When the annotators assign the rhetorical relation for a node of the tree, they provide the relation's label (drawn from the pre-defined set of relations in the annotation guidelines) and also indicate whether the spans that comprise the relation are nuclei or satellites. Figure 1 shows an example of an annotated RST tree for a response with a proficiency score of 1. This response includes three disfluencies (EDUs 3, 6, and 9), which are satellites of the corresponding repair nuclei. In addition, the response also includes an awkward Comment-Topic relation between EDU 2 and the node combining EDUs 3-11, indicated by *awkward-Comment-Topic-2*; in this multinuclear relation, the annotator judged that the second branch of the relation was awkward, which is indicated by the *2* that was appended to the relation label.

3.2 Pilot Annotation

The manual annotations were provided by two experts with prior experience in various types of data annotation on both text and speech. First, a pilot annotation was conducted to train and calibrate the annotators based on 48 training samples drawn from the TOEFL® Practice Online (TPO) product[3], which offers practice tests simulating the TOEFL iBT testing experience with authentic test questions. The training samples were selected from a TPO test form with 6 test questions and were balanced according to test question and proficiency score, i.e., 2 responses from each score level for each question.

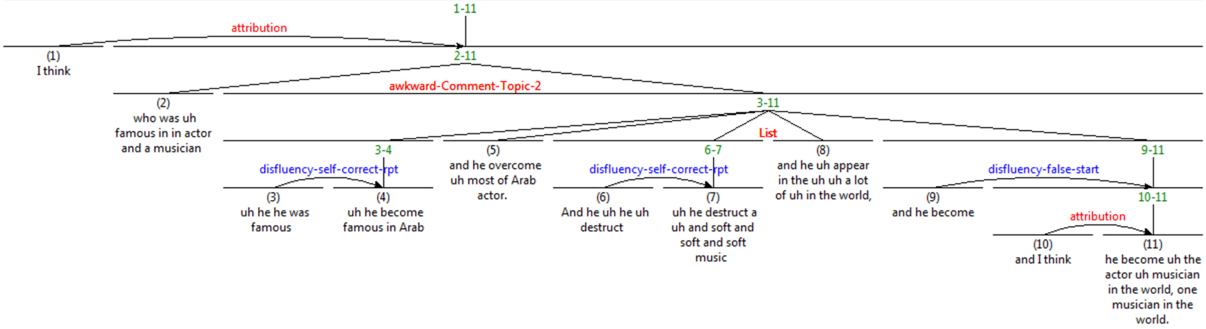
Human annotators were trained in a two-step process: 1) after a comprehensive study of the annotation guidelines described in Section 3.1, the two annotators were initially trained with 16 TPO responses (8 responses from an Independent question and 8 responses from an Integrated question) by first performing independent annotation and then resolving all disagreements through a discussion of the guidelines; 2) another round of training was conducted on an additional set of 32 TPO responses (8 responses from an Independent question and 24 responses from an Integrated question). Each annotator first annotated this set of 32 responses independently; the two annotators subsequently conducted a thorough joint review and discussion of each other's annotations in order to resolve all disagreements on this set.

In order to measure the human agreement on

[2] Downloaded from `http://www.isi.edu/licensed-sw/RSTTool/index.html`

[3] `https://toeflpractice.ets.org/`

Figure 1: Example of an annotated RST tree on a response with a proficiency score of 1.


1-11
attribution
(1)
I think
2-11
awkward-Comment-Topic-2
(2)
who was uh famous in in actor and a musician
3-11
List
3-4
disfluency-self-correct-rpt
(3)
uh he he was famous
(4)
uh he become famous in Arab
(5)
and he overcome uh most of Arab actor.
6-7
disfluency-self-correct-rpt
(6)
And he uh he uh destruct
(7)
uh he destruct a uh and soft and soft and soft music
(8)
and he uh appear in the uh uh a lot of uh in the world,
9-11
disfluency-false-start
(9)
and he become
10-11
attribution
(10)
and I think
(11)
he become uh the actor uh musician in the world, one musician in the world.

the EDU segmentation task, we first converted the segmentation sequences into 0/1 sequences: for each word in a response, 1 is assigned if a segment boundary exists after the word; otherwise, 0 is assigned. The inter-annotator agreement rate on the EDU segmentations of the 32 pilot samples (from stage 2) was κ = 0.876. On the hierarchical tree building task, inter-annotator agreement was evaluated on the levels of Span (assignment of discourse segment), Nuclearity (assignment of nucleus vs. satellite), and Relation (assignment of rhetorical relation) using κ, as described in (Marcu et al., 1999); on the 32 samples, the κ values are 0.861, 0.769, and 0.631 for the three levels, respectively.

Table 1: Human agreement on RST annotations in terms of κ and F1-Measure.

	Span	Nuclearity	Relation
κ	0.848	0.766	0.653
F1-Measure	0.872	0.724	0.522

Table 2: The average number of awkward relations appearing in responses from each of the four proficiency score levels.

	1	2	3	4
Annotator 1	3.2	1.1	1.1	0.3
Annotator 2	2.1	1.2	0.7	0.3

3.3 Formal Annotation

For the formal annotation on the full set of 600 TOEFL spoken responses, 120 responses from 6 test questions (5 responses from each score level from each question) were selected for double annotation, and the remaining 480 responses received a single annotation.

The 120 double-annotated responses were split into two batches of equal size and the two annotators each performed EDU segmentation independently on one of the batches. Subsequently, the annotators reviewed each other's EDU segmentations and adjudicated all disagreements to obtain gold-standard EDU segmentation for the 120 responses in the double-annotation set. The average number of EDUs per response in the two batches in this set of 120 responses were 15.1 and 14.1. The annotators subsequently performed the remaining steps of RST annotation (assigning the relations, nuclearity, and hierarchical structure) independently on all 120 responses using the adjudicated EDU segmentations. Table 1 shows that the κ agreements on the three levels of Span, Nuclearity, and Relation are 0.848, 0.766, and 0.653, respectively. Besides the κ evaluation, the standard ways of F1-Measure on three levels of Span, Nuclearity, and Relation (Marcu, 2000), commonly used to evaluate the performance of RST parsers, are also reported in Table 1. The F1-measures were calculated according to each pair of trees from two annotators on the same sample and then averaged across all samples, i.e., a macroaveraged F1-measure.

The human agreement results also indicate that two annotators tend to agree better on responses from speakers with higher speaking proficiency levels. This is demonstrated by positive correlations between the F1 agreement scores and the human proficiency ratings: 0.197 for Span annotations, 0.210 for Nuclearity, and 0.188 for Relation.

In addition, we also examined the distribution of the manually identified awkward relations. As shown in Table 2, awkward points occur with higher frequency in responses with lower proficiency scores.

4 Discourse Features

The ultimate aim of this line of research is to use an RST-annotated corpus to investigate features for automatically assessing discourse structure in spontaneous non-native speech. Using the annotated discourse trees, we extracted several different features based on the distribution of relations and the structure of the trees, including the number of EDUs (n_edu), the number of relations (n_rel), the number of awkward relations (n_awk_rel), the number of rhetorical relations, i.e., relations that were neither classified as awkward nor as a disfluency (n_rhe_rel), the number of different types of rhetorical relations (n_rhe_relTypes), the percentage of rhetorical relations (perc_rhe_rel) out of all relations, the depth of the RST trees (tree_depth), and the ratio between n_edu and tree_depth (ratio_nedu_depth). Table 3 lists the Pearson correlation coefficients of these features with both the holistic proficiency scores and the discourse coherence scores and demonstrates the effectiveness of these features. The n_rhe_rel feature achieves the highest correlation of 0.691 with the holistic proficiency scores, and the normalized feature perc_rhe_rel achieves the highest correlation of 0.612 with the discourse coherence scores. It is interesting to note that RST-based discourse features generally have higher correlations with the holistic speaking proficiency scores than with the more specific discourse coherence scores. This result is somewhat unexpected, since the holistic proficiency scores are based only partially on discourse coherence and also cover other aspects of speaking proficiency, such as pronunciation, fluency, grammar, and vocabulary. One potential explanation for the higher correlations could be the difference in score range (1-4 for the holistic proficiency scores and 1-3 for the discourse scores). In addition, as described in Section 2, the data set used in this study was created using a stratified random sample with an even distribution of holistic scores (which may increase the features' correlations with holistic scores), but this constraint does not apply to the discourse coherence scores.

Table 3: Pearson correlation coefficients (r) of discourse features with both the holistic proficiency scores as well as the discourse coherence scores. For the 120 double-annotated responses, the averaged feature values were used.

Features	Proficiency	Coherence
n_edu	0.58	0.397
n_rel	0.584	0.396
n_awk_rel	-0.396	-0.509
n_rhe_rel	0.691	0.541
n_rhe_relTypes	0.64	0.557
perc_rhe_rel	0.589	0.612
tree_depth	0.365	0.25
ratio_nedu_depth	0.529	0.367

5 Conclusion and Future Work

In this study, we obtained discourse coherence annotations based on Rhetorical Structure Theory for a corpus of 600 non-native spontaneous spoken responses drawn from a standardized assessment of non-native English. The RST annotation results show that the annotators achieved similar inter-annotator agreement rates as have been reported in previous studies that investigated well-formed written text (Marcu et al., 1999). In addition, we demonstrate the potential of using features derived from these RST annotations for assessing non-native spoken English through moderately high correlations with both holistic speaking proficiency scores and discourse coherence scores; the highest performing feature when evaluated on the discourse coherence scores provided by expert raters was the percentage of rhetorical relations in the entire spoken response (perc_rhe_rel), with a correlation of 0.612.

In the future, we will continue this work by addressing the following main research questions: a) how can we develop additional effective features from the discourse trees; b) how well can an automatic discourse parser trained on the obtained annotations perform; c) how well will the proposed features perform when extracted using an automatic RST parser; d) how well will the features perform when using an automated speech recognizer (rather than human transcribers) to obtain the textual transcriptions of a spoken response.

References

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the write stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems* 18(1):32–39.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. Technical Report ISI-TR-545, ISI Technical Report.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurows. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *2nd SIGDIAL Workshop on Discourse and Dialogue*. Aalborg, Denmark, pages 1–10.

In Demirahin and Deniz Zeyrek. 2014. Annotating discourse connectives in spoken Turkish. In *The 8th Liguistic Annotation Workshop*. Dublin, Ireland, pages 105–109.

ETS. 2012. The official guide to the TOEFL® test. *Fourth Edition, McGraw-Hill* .

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland, pages 511–521.

Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. 2014. The impact of deep hierarchical discourse structures in the evaluation of text coherence. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland, pages 940–949.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pages 486–496.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse (Text)* 8(3):243–281.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

Daniel Marcu, Estibaliz Amorrortu, and Magdalena Romera. 1999. Experiments in constructing a corpus of discourse trees. In *ACL Workshop on Standards and Tools for Discourse Tagging*. College Park, Maryland, pages 48–57.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Livio Robaldo. 2008. The Penn Discourse TreeBank 2.0. In *The 6th International Conference on Language Resources and Evaluation (LREC)*. Marrakech, Morocco, pages 2961–2968.

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *The Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia, pages 1039–1046.

Ted J. M. Sanders, Wilbert P. M. Spooren, and Leo G. M. Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes* 15(1):1–35.

Svetlana Stoyanchev and Srinivas Bangalore. 2015. Discourse in customer care dialogues. Poster presented at the Workshop of Identification and Annotation of Discourse Relations in Spoken Language. Saarbrücken, Germany.

Maite Taboada and William C. Mann. 2006a. Applications of Rhetorical Structure Theory. *Discourse Studies* 8(4):567–588.

Maite Taboada and William C. Mann. 2006b. Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies* 8(3):423–459.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. Annotation of discourse relations for conversational spoken dialogs. In *The Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valetta, Malta, pages 2084–2090.

Xinhao Wang, Keelan Evanini, and Klaus Zechner. 2013. Coherence modeling for the automated assessment of spontaneous spoken responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia, pages 814–819.